Principal Component Analysis

ST 560
Fall 2014

Linear Dimension Reduction

- Describe (most of) variation in the multivariate data using a smaller set of variables
- PCA can be thought of as a form of data reduction
- We want use a fewer number of new variables which contain most variational information that is in the full data

PCA

- We create a new set of variables $Z_1, Z_2, \ldots, Z_p$, each of which is a linear combinations of $X_1, X_2, \ldots, X_p$:
  - $Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$
  - $Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$
  - $\vdots$
  - $Z_p = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$
- We further require these new variables to be uncorrelated.
- This assures us that the information in Z_2 doesn't overlap with the information in Z_1.
- Having p of these new variables does not give us any data reduction.
- We would like to choose only the first m of these ($m < p$) to focus on.

First PC

- Choose a normalized linear combination Z_1 of $X_1, X_2, \ldots, X_p$ so that it accounts for as much of the variation in the original variables as possible.
- Choose a weighting vector $\mathbf{v} = (v_1, \ldots, v_p)^T$ which maximizes variance of the standardized linear combination:

$$Z = v_1 X_1 + \cdots + v_p X_p$$

- Choose $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{v}^T \mathbf{v} = 1$ to maximize

$$\begin{aligned} Var(\mathbf{v}^T X) &= \mathbf{v}^T Var(X) \mathbf{v} \\ &= \mathbf{v}^T \Sigma \mathbf{v} \end{aligned}$$

First PC

- From the quadratic form maximization theorem,

$$Z_1 = v_{11}X_1 + \cdots + v_{1p}X_p$$

where $\mathbf{v}_1 = (v_{11}, \cdots, v_{1p})^T$ is the normalized eigenvector corresponding to the largest eigenvalue of Σ, captures the majority of the variance.

- In particular,

$$\begin{aligned} Var(Z_1) &= Var(v_{11}X_1 + \cdots + v_{1p}X_p) \\ &= \mathbf{v}_1^T \Sigma \mathbf{v}_1 \\ &= \lambda_1 \end{aligned}$$

Second PC

- Find a normalized linear combination $Z_2 = v_1 X_1 + v_2 X_2 + \ldots + v_p X_p = \mathbf{v}^T X$ that has maximum variance of all linear combinations uncorrelated with $Z_1 = \mathbf{v}_1^T X$.
- Lack of correlation:

$$
\begin{aligned}
0 &= Cov(Z_2, Z_1) \\
&= Cov(\mathbf{v}^T X, \mathbf{v}_1^T X) \\
&= \mathbf{v}^T Var(X) \mathbf{v}_1 \\
&= \mathbf{v}^T \mathbf{V} \Lambda \mathbf{V}^T \mathbf{v}_1 \\
&= \mathbf{v}^T \lambda_1 \mathbf{v}_1
\end{aligned}
$$

- Thus, Z_2 is orthogonal to Z_1 in the statistical sense (uncorrelated) and in the geometric sense (the inner product of $\mathbf{v}$ and $\mathbf{v}_1$ being zero.)

Second and subsequent PCs

- Now, we want to maximize

$$Var(\mathbf{v}^T X) = \mathbf{v}_1^T \Sigma \mathbf{v}_1 \text{ subject to } \mathbf{v}^T \mathbf{v} = 1 \text{ and } \mathbf{v}^T \mathbf{v}_1 = 0.$$

- Recall the quadratic form maximization theorem under the orthogonality constraints:

$$\begin{aligned} \max_{\substack{\mathbf{v} \in \mathbb{R}^p, \mathbf{v}^T \mathbf{v} = 1 \\ \mathbf{v}^T \mathbf{v}_1 = 0, \ldots, \mathbf{v}^T \mathbf{v}_{k-1} = 0}} \mathbf{v}^T \Sigma \mathbf{v} &= \mathbf{v}_k^T \Sigma \mathbf{v}_k \\ &= \lambda_k, \end{aligned}$$

- We get the second PC, $Z_2 = \mathbf{v}_2^T X$ (and subsequent PCs, $Z_k = \mathbf{v}_k^T X$).

Population PCA

Let the p-variable vector $\mathbf{X} = (X_1, \ldots, X_p)^\intercal$ have the covariance matrix $\Sigma = \mathbf{V}\Lambda\mathbf{V}^T$ with

- $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \ldots \geq \lambda_p$
- $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ orthonormal
- $\Sigma\mathbf{v}_i = \lambda_i\mathbf{v}_i$.

Then,

- The kth eigenvector $\mathbf{v}_k$ is the kth PC direction vector.
- The kth eigenvalue λ_k is the variance explained by the kth principal component score (random variable), $Z_k = \mathbf{v}_k^T\mathbf{X}$: $Var(Z_k) = \lambda_k, \; k = 1, \ldots, p$
- $Cov(Z_i, Z_j) = \mathbf{v}_i^T\Sigma\mathbf{v}_j = 0, \; i \neq j$.

PC in practice?

The first part of this chapter deals with PC from a population covariance. In practice, the covariance matrix is unknown. For a random sample $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]_{p \times n}$ from a population, the sample PCA sequentially finds orthogonal directions of maximal (projected) sample variance.

- Define the centered data matrix

$$\tilde{\mathbf{X}} = [\mathbf{X}_1 - \bar{\mathbf{X}}, \mathbf{X}_2 - \bar{\mathbf{X}}, \cdots, \mathbf{X}_n - \bar{\mathbf{X}}].$$

- The sample variance-covariance matrix is

$$\mathbf{S}_n = \frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T.$$

- Eigen-decomposition of $\mathbf{S}_n = \hat{\mathbf{V}}\hat{\Lambda}\hat{\mathbf{V}}^\top$ leads to the sample PC directions $(\hat{\mathbf{v}}_k)$ and the variance of the kth sample scores $(\hat{\lambda}_k)$.

Sample PCA

From the quadratic form maximization theorem, we can verify that

$$\hat{\mathbf{v}}_k = \arg \max_{\substack{\mathbf{v} \in \mathbb{R}^p, \mathbf{v}^T\mathbf{v} = 1 \\ \mathbf{v}^T\hat{\mathbf{v}}_1 = 0, \ldots, \mathbf{v}^T\hat{\mathbf{v}}_{k-1} = 0}} \tilde{V}ar(\mathbf{v}^{\mathsf{T}}\mathbf{X}),$$

where $\tilde{V}ar$ denotes the sample covariance. Thus,

- $\hat{\mathbf{v}}_k$ is the kth sample PC direction vectors and is the vector of the kth loadings.
- $\mathbf{z}_k = \left(\hat{\mathbf{v}}_k^{\mathsf{T}}\tilde{(}\mathbf{X}_i - \bar{\mathbf{X}})\right)_{i=1}^n = \hat{\mathbf{v}}_k^{\mathsf{T}}\tilde{\mathbf{X}}$ is the kth score vector.
- $\hat{\lambda}_k =$ the sample variance of $\mathbf{z}_k$

SVD and PCA

Consider Singular Value Decomposition of the centered data matrix: $\tilde{\mathbf{X}}^T = \mathbf{UDV}^T$. Then,

$$\begin{aligned}
\mathbf{S}_n &= \frac{1}{n-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T \\
&= \frac{1}{n-1}(\mathbf{UDV}^T)^T(\mathbf{UDV}^T) \\
&= \mathbf{V}\text{diag}(\frac{1}{n-1}d_k^2)\mathbf{V}^T.
\end{aligned}$$

Thus,

- PC directions: (1) the right singular vectors of the centered data matrix of size $n \times p$ or (2) the eigenvectors of the sample covariance matrix $\mathbf{S}$.
- Variance of PC scores: (1) the squared singular values of the (centered) data matrix/$(n-1)$ or (2) the eigenvalues of the sample covariance matrix $\mathbf{S}$.

Computation of PCA

PCA is computed

1. using eigenvalue-eigenvector decomposition of $\mathbf{S}$ or
2. using the singular value decomposition of $\tilde{\mathbf{X}}_{p\times n} = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)_{i=1}^{n}$

From $\mathbf{S} = \mathbf{V}\Lambda\mathbf{V}^T$

1. PC directions : $\mathbf{v}_k$ (eigenvectors)
2. the kth sample pc (or score vectors): $\mathbf{z}_k = \hat{\mathbf{v}}_k^{\top}\tilde{\mathbf{X}}$
3. Variance of kth PC score: λ_k (eigenvalues)

From SVD of $\tilde{\mathbf{X}}^{\top} = \mathbf{UDV}^T$

1. PC directions : $\mathbf{v}_k$ (right singular vectors)
2. the kth sample pc (or score vectors): $\mathbf{z}_k = \hat{\mathbf{v}}_k^{\top}\tilde{\mathbf{X}}$
3. Variance of k-th PC score: $\frac{1}{n-1}d_k^2$ (singular values^2)

Eigen-expansion

Eigen-expansion of the data matrix:

$$\tilde{\mathbf{X}}_{p\times n} = (\sum_{i=1}^{p} \mathbf{v}_i \mathbf{v}_i^{\mathsf{T}})\tilde{\mathbf{X}}$$

Then, the raw data matrix can be written as

$$\begin{aligned}
\mathbf{X}_{p\times n} &= \bar{\mathbf{X}} + \tilde{\mathbf{X}} \\
&= \bar{\mathbf{X}} + \sum_{i=1}^{p} \mathbf{v}_i(\mathbf{v}_i^{\mathsf{T}}\tilde{\mathbf{X}}) \\
&= \bar{\mathbf{X}} + \sum_{i=1}^{p} \mathbf{v}_i \mathbf{z}_i
\end{aligned}$$

$\mathbf{z}_i$: scores, observed i-th PC...

$\mathbf{v}_i$: loadings, eigenvectors, PC direction vectors...

Reduced Rank Representation:

Reconstruct using only the first few terms (assuming decreasing eigenvalues)

$$\mathbf{X}_m \approx \bar{\mathbf{X}} + \sum_{i=1}^{m} \mathbf{v}_i \mathbf{z}_i$$

gives rank m approximation of data

- The larger m, the better approximation by $\mathbf{X}_m$
- The smaller m, the more succinct dimension reduction of $\mathbf{X}$

Renaming

- Statistics: Principal Component Analysis (PCA)
- Social Sciences: Factor Analysis (PCA is a subset)
- Probability/ Electrical Eng: Karhunen - Loeve expansion
- Applied Mathematics: Proper Orthogonal Decomposition (POD)
- Geo-Sciences: Empirical Orthogonal Functions (EOF)

Covariance vs Correlation

- Often the variables in the raw data set are very different in their scales, variabilities, etc.
- Basing the PCA on the covariance matrix would lead to variables with large variances dominating the most important principal components
- Also, changing the units of measurements would change the PCA solution.
- For this reason, it is often preferred to base the PCA solution on the eigenvectors and eigenvalues of the correlation matrix rather than the covariance matrix.
- This is equivalent to initially standardizing all variables and then performing the PCA base on the correlation matrix.

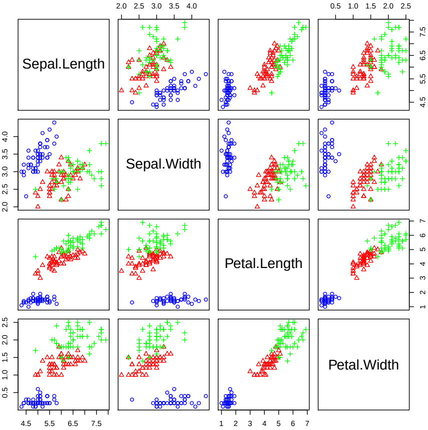
Example 1: Iris Data

- This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day.
- The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
- Predicted attribute: class of iris plant.
- Number of Instances: 150 (50 in each of three classes)
- Number of Attributes: 4 numeric, predictive attributes and the class (Iris Setosa, Iris Versicolour,Iris Virginica)

Iris Data

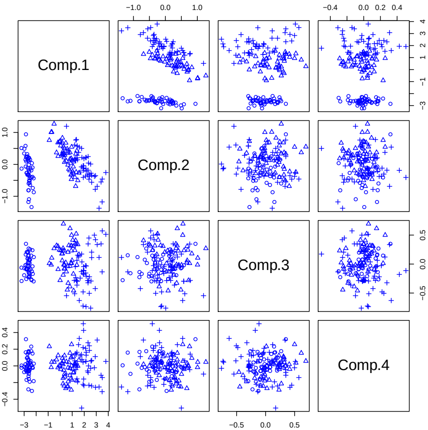
- Attribute Information:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm

Iris Data: scatter plot



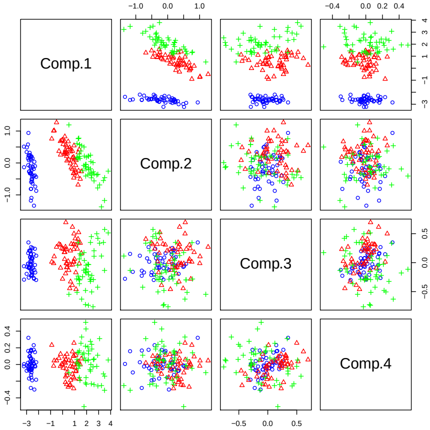
(i, j)th frame: scatter plot of $(\mathbf{x}_i, \mathbf{x}_j)$ (Iris Setosa(b), Iris Versicolour (r), Iris Virginica(g))

Iris Data: PCA scat plot



(i, j)th frame: scatter plot of $(\mathbf{z}_i, \mathbf{z}_j)$

Iris Data: PCA scat plot



(i, j)th frame: scatter plot of $(\mathbf{z}_i, \mathbf{z}_j)$
(Iris Setosa(b), Iris Versicolour (r), Iris Virginica(g))

Iris Data: how many components to keep?

The criterion for PCA is a high variance in the principal components. The question involves "how much the PCs explain the variance present in the whole data?"

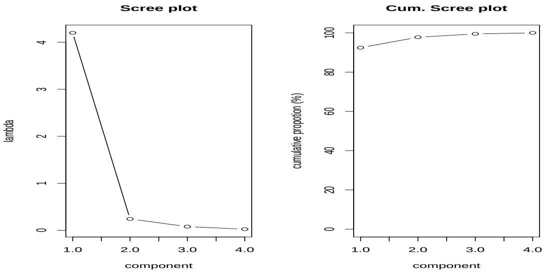
- Note that Variance of kth PC score is λ_k (kth eigenvalue of $\mathbf{S}$).
- Total variance in the whole data is the same as the total variance explained by all PCs:

$$\sum_{k=1}^{p} Var(k\text{th PC score}) = \sum_{i=1}^{p} \lambda_k$$
$$= \text{trace}(\mathbf{S}) = \sum_{k=1}^{p} Var(X_k)$$

- Variance explained by the first k PCs is $\lambda_1 + \ldots + \lambda_k$.

Iris Data: scree plot

In scree plot (k, λ_k), look for an elbow. In cumulative scree plot, proportion of variance explained $\left(k, \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}\right)$, use 90% as a cutoff.



Which variables are most responsible for the PCs?

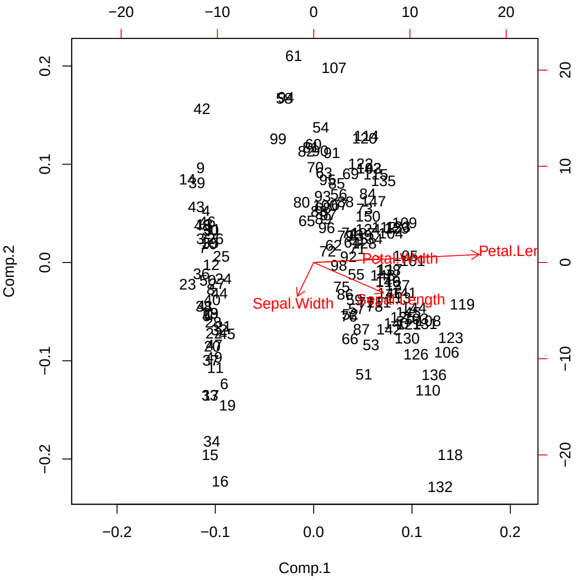
- Loadings of PC directions.
- Biplot- scatter plot of PC1 and PC2 scores, overlaid with p vector each representing the loadings of the first two PC directions.

Iris Data: PC Direction

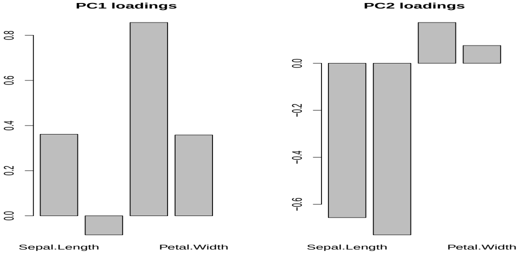
In the Iris data, the PC loadings are

```
Loadings:
             Comp.1 Comp.2 Comp.3 Comp.4
Sepal.Length  0.361 -0.657  0.582  0.315
Sepal.Width         -0.730 -0.598 -0.320
Petal.Length  0.857  0.173        -0.480
Petal.Width   0.358        -0.546  0.754
```

Iris Data: biplot

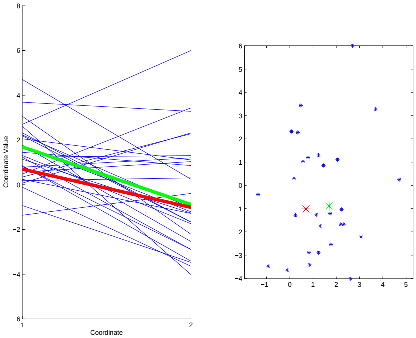


Iris Data: PCA



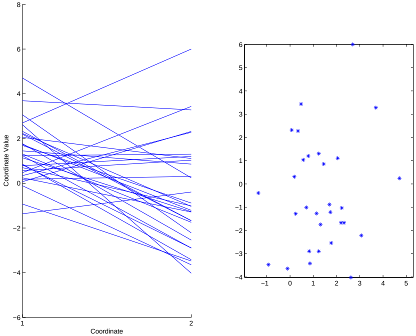
- The 1st PC accounts for 92% of the total variation.
- From the 1st PC loadings: 1st PC is the weighted average of sepal length, petal length and petal width.

Example 2: PCA for curve data



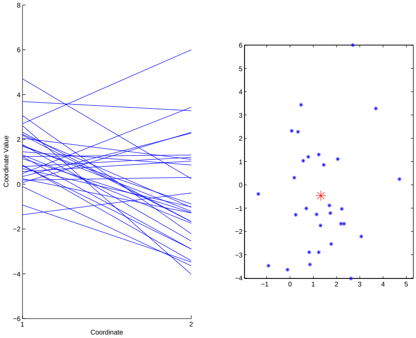
Data Points (Curves) are columns of data matrix, **X**.
Two data points are highlighted.

PCA for curve data



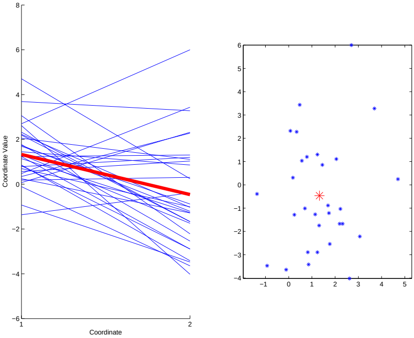
Data Points (Curves) are columns of data matrix, $\mathbf{X}$.

PCA for curve data



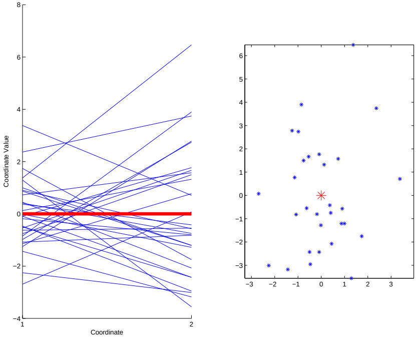
Sample mean in 2-d space

PCA for curve data



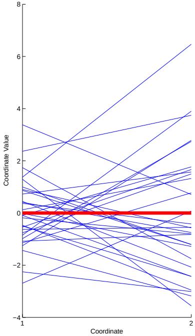
Sample mean in the curve space

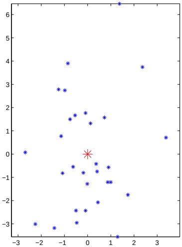
PCA for curve data



Mean Centered Data

PCA for curve data



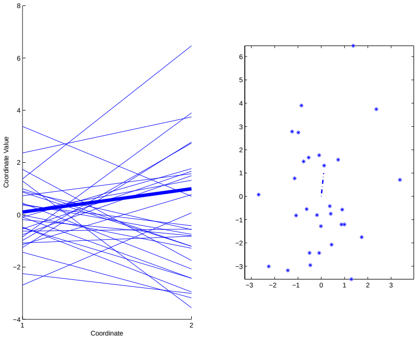


PCA with Mean Centered Data

$$\hat{\mathbf{V}} = \begin{pmatrix} 0.1041 & -0.9946 \\ 0.9946 & 0.1041 \end{pmatrix}$$

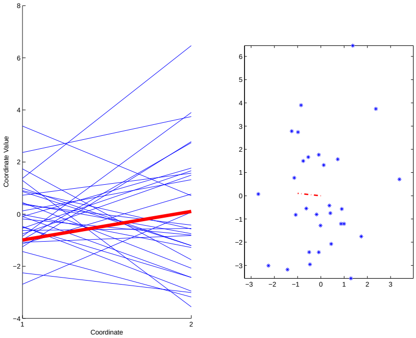
$$\hat{\Lambda} = \begin{pmatrix} 5.7939 & 0 \\ 0 & 1.6480 \end{pmatrix}$$

PCA for curve data



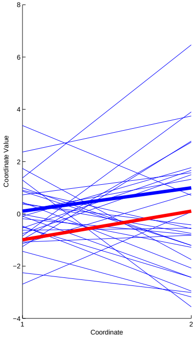
PC 1 direction $\hat{\mathbf{v}}_1 = \binom{0.1041}{0.9946}$

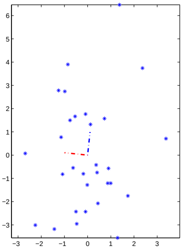
PCA for curve data



PC 2 direction $\hat{\mathbf{v}}_2 = \begin{pmatrix} -0.9946 \\ 0.1041 \end{pmatrix}$

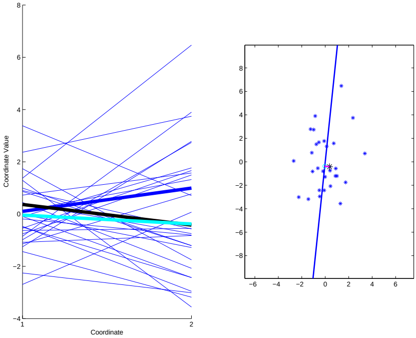
PCA for curve data





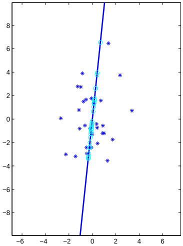
PC 1 and 2 directions

PCA for curve data



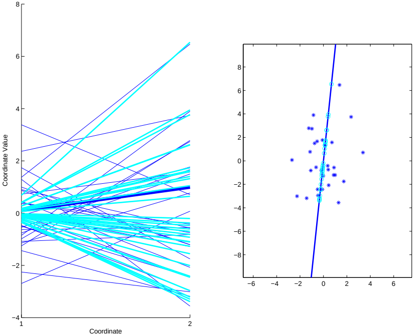
Projection of one data vector onto PC1 direction

PCA for curve data



Projection of data vectors onto PC1 direction

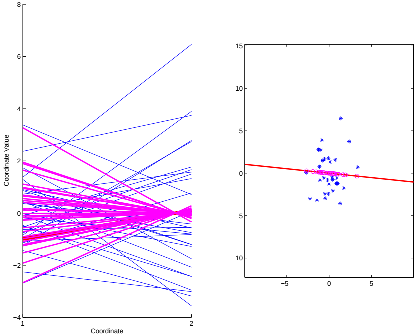
PCA for curve data



Projection of data vectors onto PC1 direction

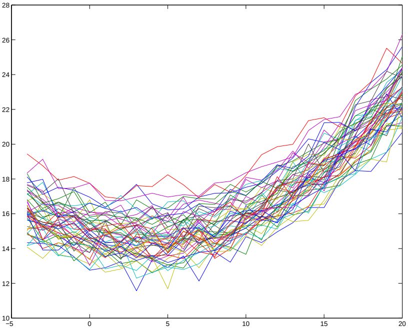
$\hat{\lambda}_1 = 5.7939$

PCA for curve data



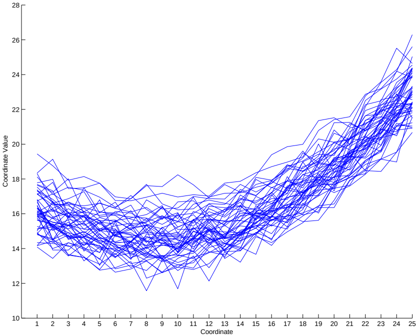
Projection of data vectors onto PC2 direction
$\hat{\lambda}_2 = 1.6480$

PCA for curve data



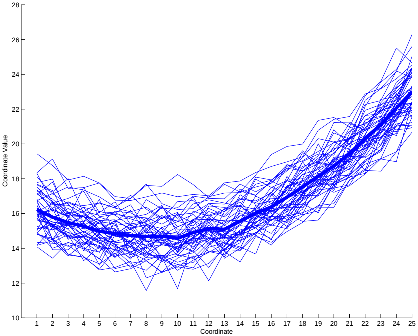
$n = 50$, $d = 25$ grid points curve data

PCA for curve data



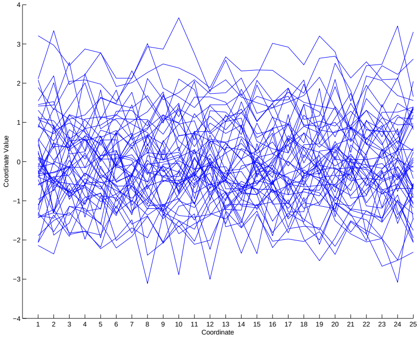
$n = 50$, $d = 25$ grid points curve data

PCA for curve data



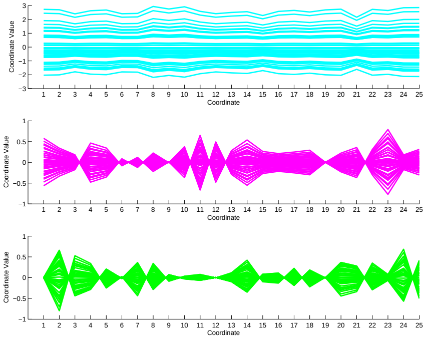
sample mean

PCA for curve data



mean centered data

PCA for curve data



PC1 - PC3 projections
$\hat{\lambda}_1 = 40.5, \hat{\lambda}_2 = .8, \hat{\lambda}_3 = .7$

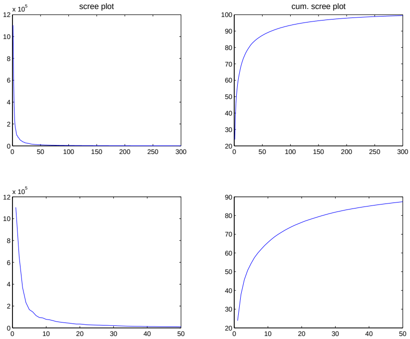
Example 3: Olivetti Faces data

- Obtained from
  `http://www.cs.nyu.edu/~roweis/data.html`
- Grayscale faces 8 bit [0-255], 10 images of 40 different people.
- $n = 400$ total images of size 64×64.

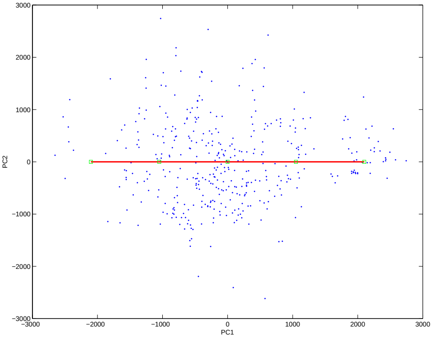
Images as data

- An image is a matrix-valued datum.
- For Olivetti Faces data, the matrix is of size 64×64, with each pixel having values between 0 and 255.
- The matrix corresponding to each observation is vectorized by stacking each column into one long vector of size $p = 64 \times 64 = 4096$.
- So, my data matrix $\mathbf{X}$ is of size 400×4096. Now, PCA is applied to this data matrix.

Face data: scree plot

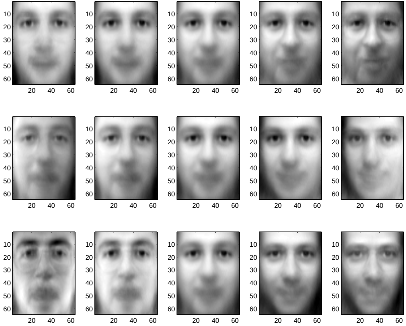


Face data: scatter plot



What are the loadings?

Face data: marching along the first 3 PC directions



- PC1: lighter to darker face
- PC2: masculin to feminin face
- PC3: rectangle to oval face, presence of eyeglasses

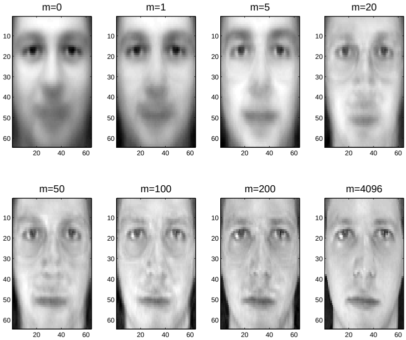
PCA approximation

Approximation to the original data matrix:

$$\hat{\mathbf{x}}_i \approx \bar{\mathbf{x}} + \sum_{j=1}^{m} \mathbf{v}_j (\mathbf{y}_j)_i$$

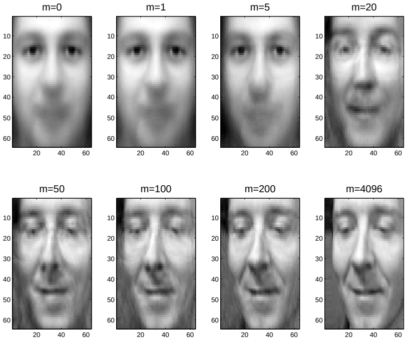
PCA approximation

Observation index $i = 1$



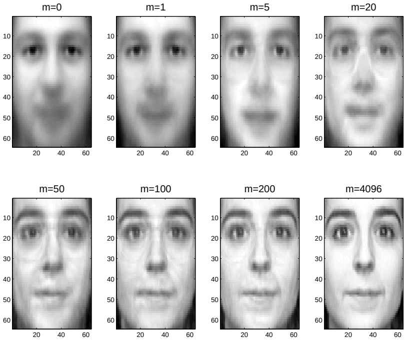
PCA approximation

Observation index $i = 25$



PCA approximation

Observation index $i = 100$



Face data PCA approximation

- Human eyes require > 50 PCs to see resemblance between $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$.
- Variance explained by 50 PCs is about 90 % of total variance.
- Reconstruction by PCA most useful when
  - each datum is visually represented (rather than being just numbers)
  - data objects are images, shapes, functions.

PC in Regression

- Predict a real-valued output Y using a set of covariates $X = (X_1, \ldots, X_p)$.
- Linear model assumes the regression function $E(Y|X)$ is linear; $E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$.
- Assume

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i,$$

where $\epsilon_i \overset{i.i.d.}{\sim} (0, \sigma^2)$ for $i = 1, \ldots, n$.

Least Squares method

- Least Squares method chooses β which minimizes the residual sum of squares:

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} X_{ji}\beta_j)^2$$

- Using matrix notation,
  - $\hat{\beta} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$ and
  - $\hat{\mathbf{y}} = \mathbf{X}^{\mathrm{T}}\hat{\beta}$,

  where $\mathbf{X}$ is the $(p+1) \times n$ design matrix with the 1's on the first row and $\beta = (\beta_0, \ldots, \beta_p)^T$.

Connecting LS estimate with PCA

- From now on, assume that data vector **y** and **X** are centered. In particular, **X** doesn't have 1's on its first row and we will fit linear model with no intercept subsequently.
- Note that for any orthonormal matrix **U**

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}^{\mathsf{T}}\beta + \epsilon \\
&= \underbrace{\mathbf{X}^{\mathsf{T}}\mathbf{U}}\,\underbrace{\mathbf{U}^{T}\beta} + \epsilon \\
&= \mathbf{Z}^{\mathsf{T}}\gamma + \epsilon
\end{aligned}
$$

- PCR chooses PC directions for **U** and PC scores for **Z**.

Why PCR?

- Least Squares estimate based on original input variables:

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y} \\ &= (\mathbf{V}\Lambda\mathbf{V}^\top)^{-1}\mathbf{X}\mathbf{y} \\ &= \mathbf{V}\Lambda^{-1}\mathbf{V}^T\mathbf{X}\mathbf{y}\end{aligned}$$

Multicollinearity

- What is (perfect) multicollinearity?
  - the p-explanatory variables are not linearly independent
  - at least one vector in the set can be written as a linear combination of the other vectors
  - the p-explanatory variables live in the q-dimensional subspace of p-dimensional space
  - there exists zero eigenvalues of Σ
- In reality, when multicollinearity exists in the data, we observe very small sample eigenvalues.

Multicollinearity in multiple regression

- Why this can be an issue?
- Observation 1:

$$\begin{aligned}\hat{\beta} &= \mathbf{V}\Lambda^{-1}\mathbf{V}^T\mathbf{X}\mathbf{y} \\ &= \sum_{j=1}^{p}(\frac{1}{\lambda_j}\mathbf{v}_j\mathbf{v}_j^T)\mathbf{X}\mathbf{y}\end{aligned}$$

- With small eigenvalues, the inversion of matrix is numerically unstable.

Multicollinearity in multiple regression

- Observation 2: Under the usual iid error assumption,

$$E(\hat{\beta}) = \beta$$

but

$$\begin{aligned} Var(\hat{\beta}) &= (\mathbf{X}\mathbf{X}^T)^{-1}\sigma^2 \\ &= \sum_{j=1}^{p} \frac{1}{\lambda_j}\mathbf{v}_j\mathbf{v}_j^T\sigma^2 \end{aligned}$$

- Change in design matrix can change the estimate drastically.

What this means in PCs

- When multicollinearity exists, it appears as PCs with very small variance, hence very large values of $\frac{1}{\lambda_k}$.
- Any predictor variable having moderate or large coefficients in any of the PCs associated with very small eigenvalues will have a very large variance.
- How to reduce this effect?

Principal Component Regression

- PC scores: $\mathbf{z}_i = \mathbf{v}_i^{\mathsf{T}}\mathbf{X}$, $i = 1, \ldots, p$.
- Let $\mathbf{Z}_k = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_k \end{bmatrix}_{k \times n}$ be the design matrix with the first k PCs.
- PCR chooses the first $k (< p)$ PCs in the regression analysis:

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\beta + \epsilon \\
&\approx \mathbf{Z}_k\gamma_k + \epsilon_k
\end{aligned}
$$

- With the first kPC, we get $\tilde{\gamma}_k = (\mathbf{Z}_k\mathbf{Z}_k^T)^{-1}\mathbf{Z}_k\mathbf{y}$, coefficient estimates of the PC scores.
- For the original variables, we get

$$
\tilde{\beta}_k = \mathbf{V}_k\tilde{\gamma}_k
$$

Principal Component Regression

- What do we gain?

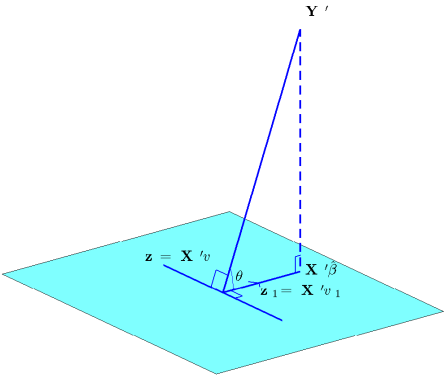
$$\begin{aligned} Var(\tilde{\beta}_k) &= Var(\mathbf{V}_k \tilde{\gamma}_k) \\ &= \mathbf{V}_k (\mathbf{Z}_k \mathbf{Z}_k^T)^{-1} \mathbf{Z}_k Var(\mathbf{y}) \mathbf{Z}_k^T (\mathbf{Z}_k \mathbf{Z}_k^T)^{-1} \mathbf{V}_k^T \\ &= \mathbf{V}_k (\mathrm{diag}(\lambda_1, \ldots, \lambda_k))^{-1} \mathbf{V}_k^T \sigma^2 \\ &= \sum_{j=1}^{k} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j^T \sigma^2 \end{aligned}$$

- What do we lose?

$$\begin{aligned} E(\tilde{\beta}_k) &= \mathbf{V}_k \underbrace{(\mathbf{Z}_k \mathbf{Z}_k^T)^{-1}} \underbrace{\mathbf{Z}_k} \underbrace{E(\mathbf{y})} \\ &= \mathbf{V}_k \overbrace{\Lambda_k^{-1}} \overbrace{\mathbf{V}_k^T \mathbf{X}} \overbrace{\mathbf{X}^T \beta} \\ &\neq \beta \end{aligned}$$

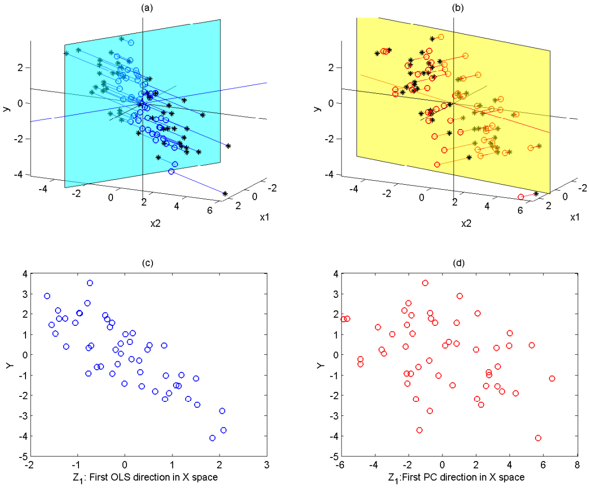
- Decrease in variance for the estimator $\tilde{\beta}_k$ is achieved at the expense of introducing bias in to the estimator $\tilde{\beta}_k$

OLS vs PCR



$\mathbf{Y}'$
$\mathbf{z} = \mathbf{X}'v$
θ
$\mathbf{X}'\hat{\beta}$
$\mathbf{z}_1 = \mathbf{X}'v_1$

OLS vs PCR



(a)
(b)
x2
x1
y
(c)
(d)
Y
Z_1: First OLS direction in X space
Z_1:First PC direction in X space

Boston Housing Example

- X_1: per capita crime rate,
- X_2: proportion of residential land zoned for large lots,
- X_3: proportion of nonretail business acres,
- X_4: Charles River (1 if tract bounds river, 0 otherwise),
- X_5 : nitric oxides concentration,
- X_6 : average number of rooms per dwelling,
- X_7 : proportion of owner-occupied units built prior to 1940,
- X_8 : weighted distances to five Boston employment centers,
- X_9 : index of accessibility to radial highways,
- X_{10} : full-value property to radial highways,
- X_{11} : pupil/teacher ratio,
- X_{12} : $1000(B - 0.63)^2 I(B < 0.63)$ where B is the proportion of African American,
- X_{13} : % lower status of the population,
- y: median value of owner-occupied homes in \$1000.

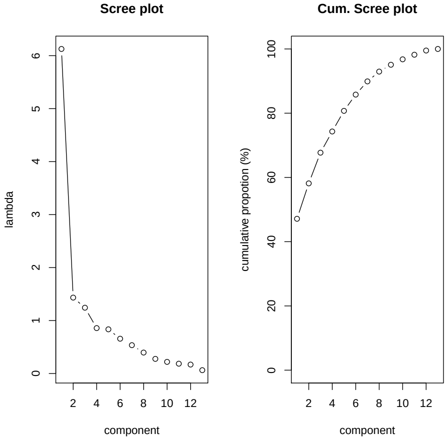
Data

- Do centering and scaling of the data so that $Var(X) = Corr(X)$.
- Do PCA and regression with the scaled data.

Correlation Matrix R

1.0000	-0.2005	0.4066	-0.0559	0.4210	-0.2192	0.3527	-0.3797	0.6255	0.5828	0.2
-0.2005	1.0000	-0.5338	-0.0427	-0.5166	0.3120	-0.5695	0.6644	-0.3119	-0.3146	-0
0.4066	-0.5338	1.0000	0.0629	0.7637	-0.3917	0.6448	-0.7080	0.5951	0.7208	0
-0.0559	-0.0427	0.0629	1.0000	0.0912	0.0913	0.0865	-0.0992	-0.0074	-0.0356	-0
0.4210	-0.5166	0.7637	0.0912	1.0000	-0.3022	0.7315	-0.7692	0.6114	0.6680	0
-0.2192	0.3120	-0.3917	0.0913	-0.3022	1.0000	-0.2403	0.2052	-0.2098	-0.2920	-0
0.3527	-0.5695	0.6448	0.0865	0.7315	-0.2403	1.0000	-0.7479	0.4560	0.5065	0
-0.3797	0.6644	-0.7080	-0.0992	-0.7692	0.2052	-0.7479	1.0000	-0.4946	-0.5344	-0
0.6255	-0.3119	0.5951	-0.0074	0.6114	-0.2098	0.4560	-0.4946	1.0000	0.9102	0
0.5828	-0.3146	0.7208	-0.0356	0.6680	-0.2920	0.5065	-0.5344	0.9102	1.0000	0
0.2899	-0.3917	0.3832	-0.1215	0.1889	-0.3555	0.2615	-0.2325	0.4647	0.4609	1
-0.3851	0.1755	-0.3570	0.0488	-0.3801	0.1281	-0.2735	0.2915	-0.4444	-0.4418	-0
0.4556	-0.4130	0.6038	-0.0539	0.5909	-0.6138	0.6023	-0.4970	0.4887	0.5440	0

PCA: Scree plot



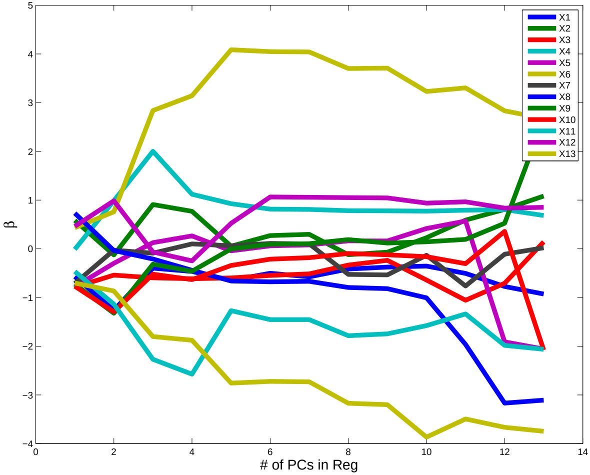
PCA loadings

Loadings:

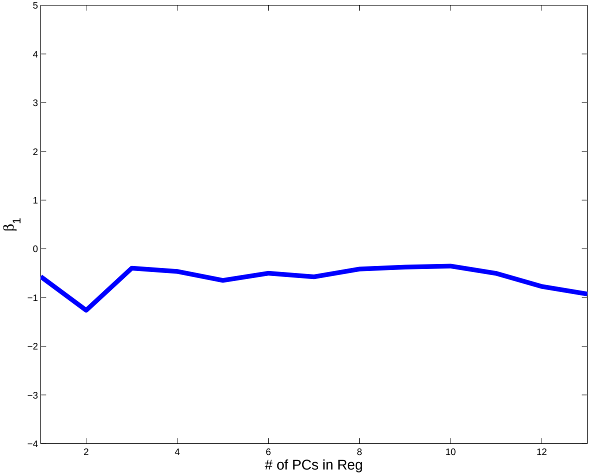
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13
crim	0.251	0.315	-0.247			0.220	0.778	0.153	0.260		-0.110		
zn	-0.256	0.323	-0.296	0.129	0.321	0.323	-0.275	-0.403	0.358	-0.268	0.263		
indus	0.347	-0.112					-0.340	0.174	0.644	0.364	-0.303	-0.113	-0.251
chas		-0.455	-0.290	0.816		-0.167							
nox	0.343	-0.219	-0.121	-0.128	0.137	0.153	-0.200			-0.231	0.111	0.804	
rm	-0.189	-0.149	-0.594	-0.281	-0.423			-0.327		0.431		0.153	
age	0.314	-0.312		-0.175			0.116	-0.601		-0.363	-0.459	-0.212	
dis	-0.322	0.349		0.215			-0.104	-0.122	-0.153	0.171	-0.696	0.391	
rad	0.320	0.272	-0.287	0.132	-0.204	0.143	-0.138		-0.471			-0.107	-0.633
tax	0.338	0.239	-0.221	0.103	-0.130	0.193	-0.315		-0.177		-0.105	-0.215	0.720
ptratio	0.205	0.306	0.323	0.283	-0.584	-0.273		-0.318	0.254	-0.153	0.175	0.210	
black	-0.203	-0.239	0.300	0.168	-0.346	0.803							
lstat	0.310		0.267		0.395			-0.424	-0.195	0.601	0.271		

- The 13th PC $\approx -0.2X_3 - 0.6X_9 + 0.7X_{10}$ has a small variance.
- This means that the 13th PC $\propto -X_3 - 3X_9 + 3X_{10}$ is near constant.

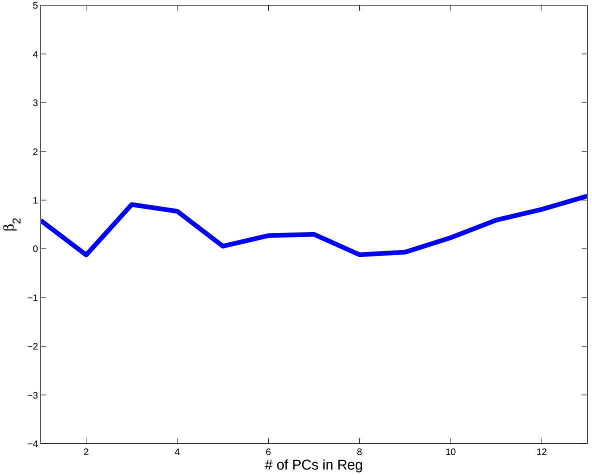
How many PCs? Coefficient changes



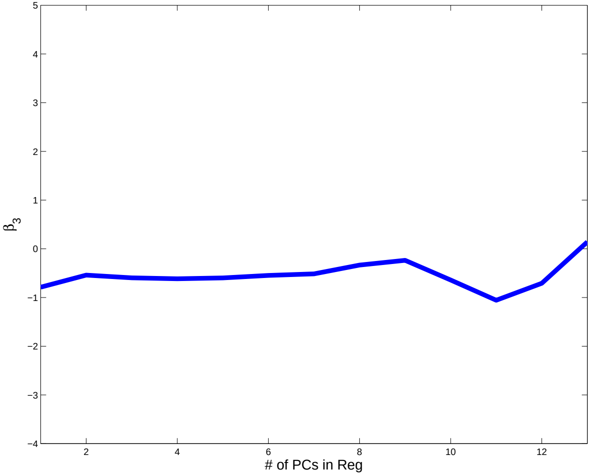
PC Regression coefficient changes: X_1



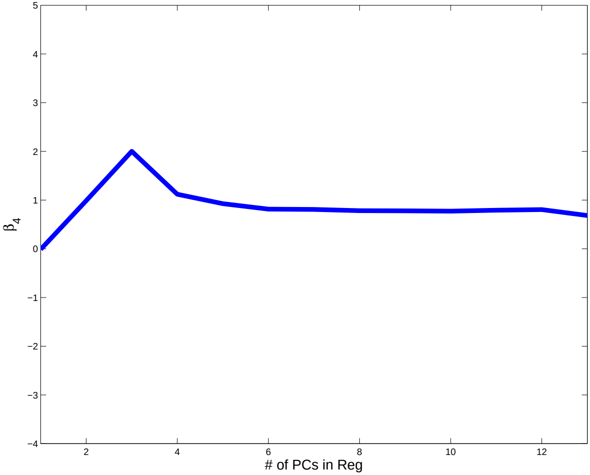
PC Regression coefficient changes: X_2



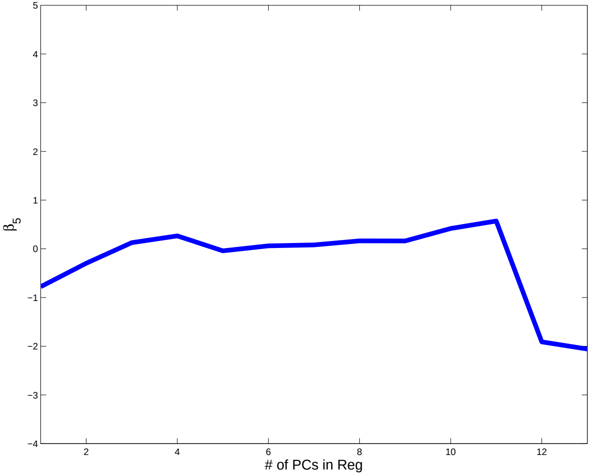
PC Regression coefficient changes: X_3



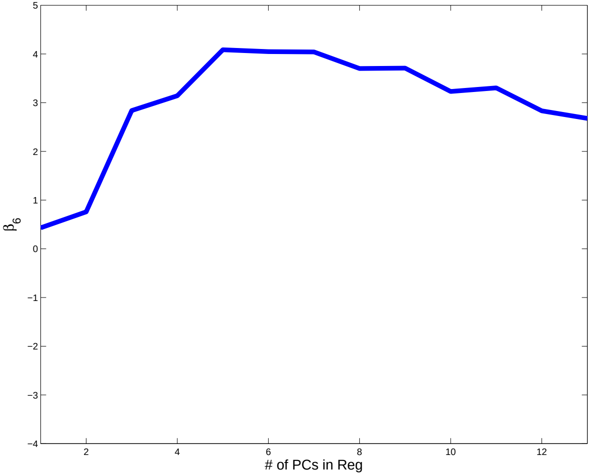
PC Regression coefficient changes: X_4



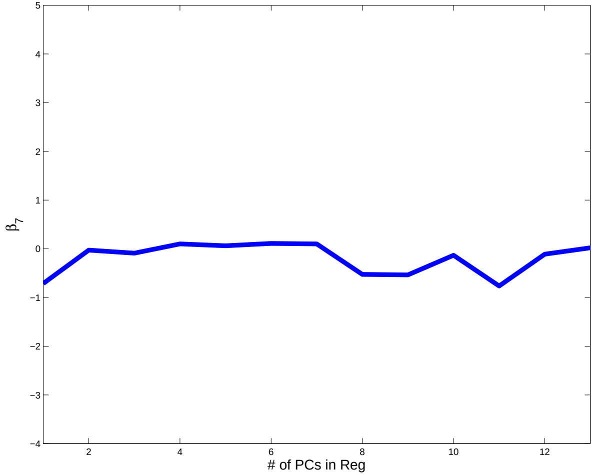
PC Regression coefficient changes: X_5



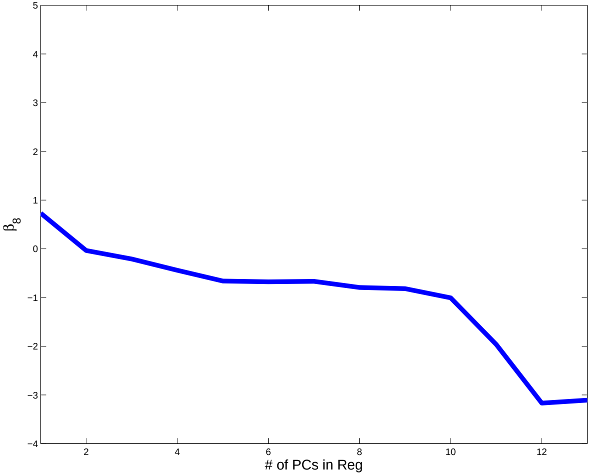
PC Regression coefficient changes: X_6



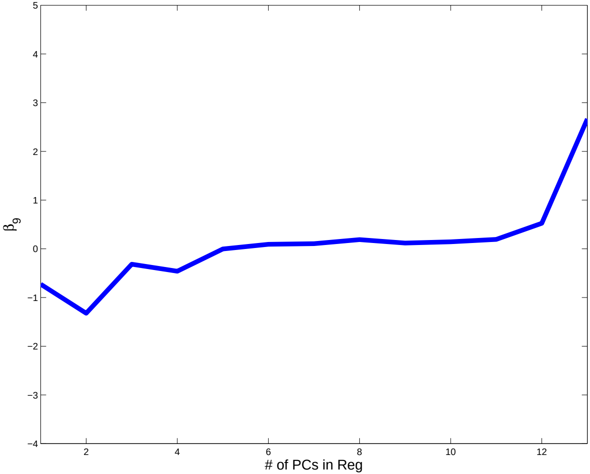
PC Regression coefficient changes: X_7



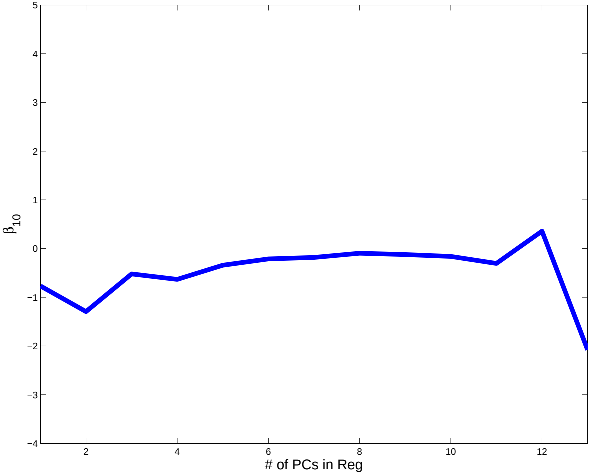
PC Regression coefficient changes: X_8



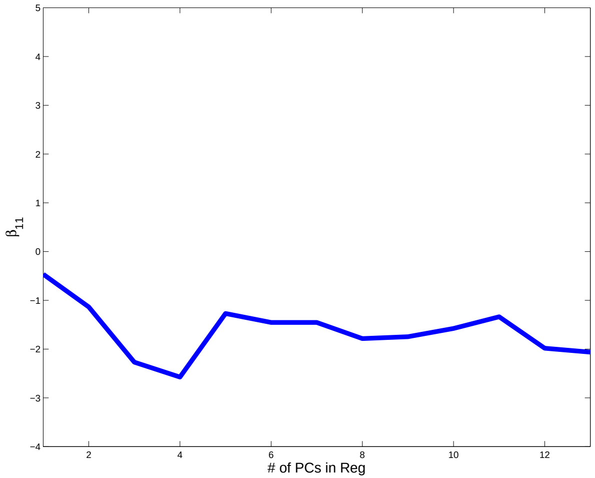
PC Regression coefficient changes: X_9



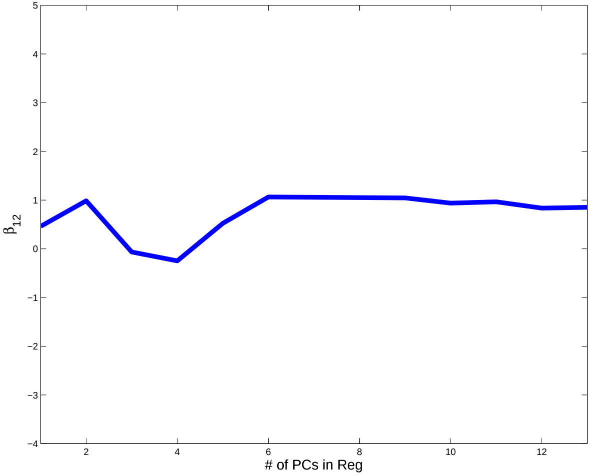
PC Regression coefficient changes: X_{10}



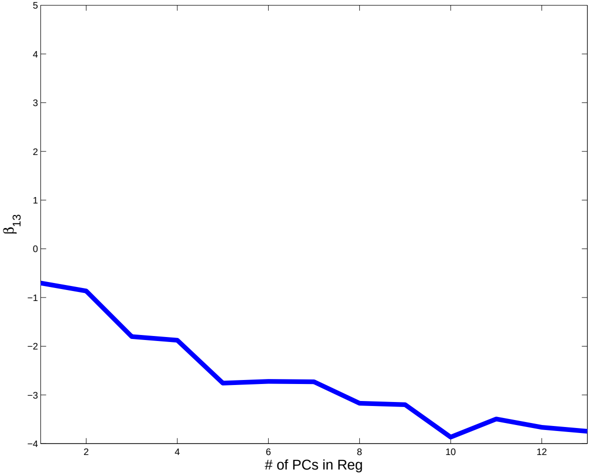
PC Regression coefficient changes: X_{11}



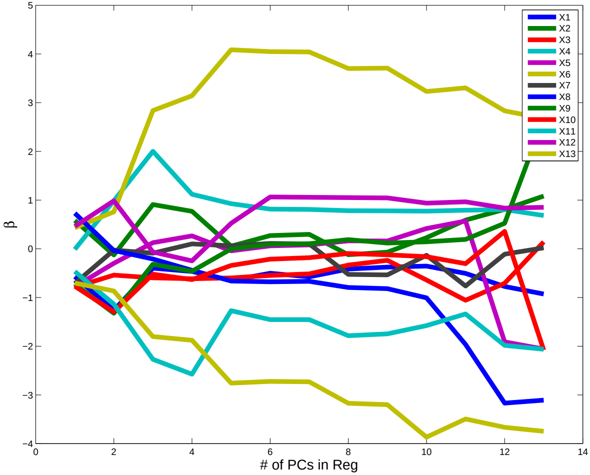
PC Regression coefficient changes: X_{12}



PC Regression coefficient changes: X_{13}



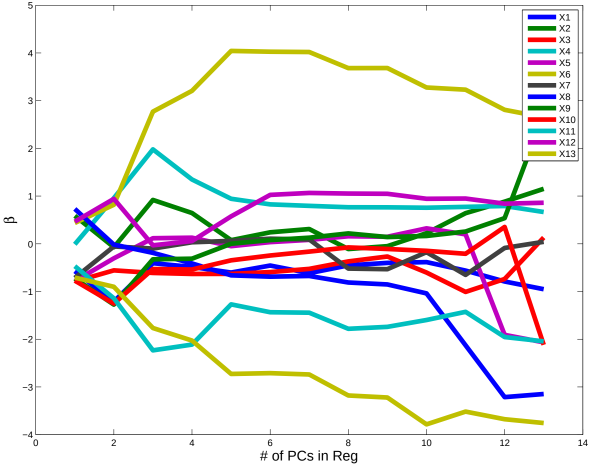
PC Regression coefficient



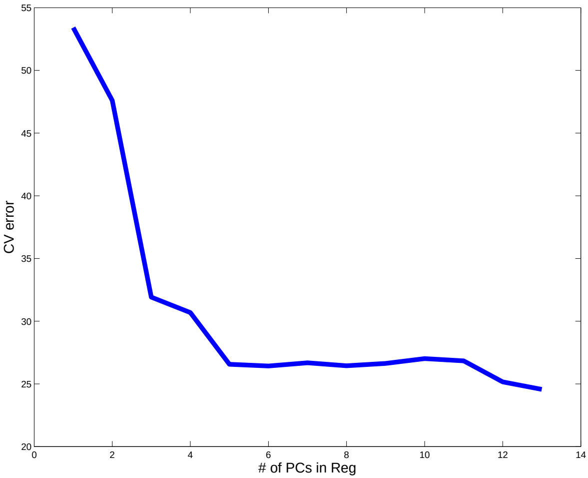
How many PCs? Predictive performance

- One possibility: Choose k based on the predictive ability.
- Randomly split the data into two part: training (380) vs test(126).
- Fit the model on the training data using the first k PCs.
- Predict y values for the test set, $\hat{y}$.
- Study the prediction error $y - \hat{y}$ in the test set.
- Do the random splitting over 50 times.
- This technique is called Cross-Validation.

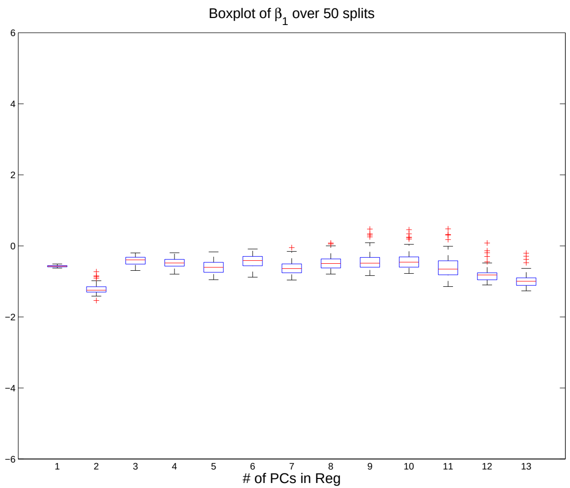
PC Regression coefficient mean over 50 random splits



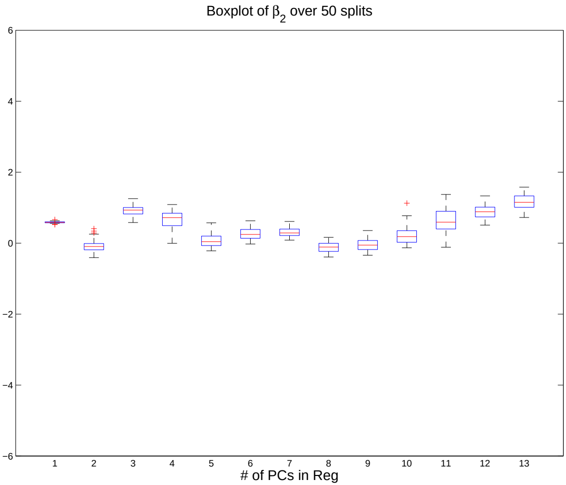
Cross-Validation



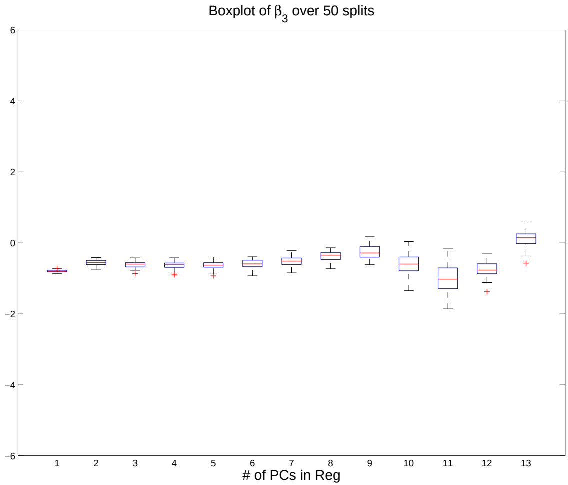
Coefficient variability over 50 random splits



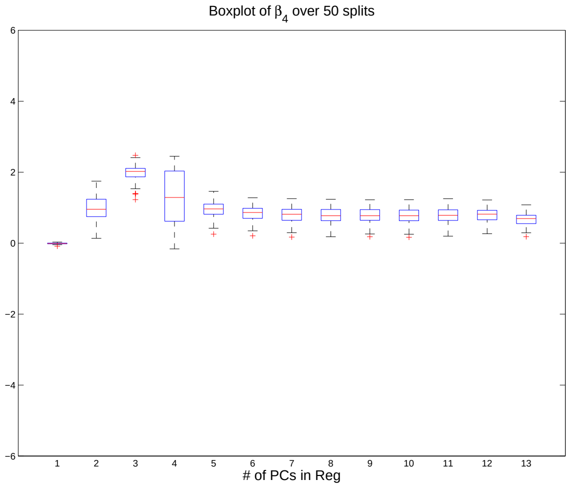
Coefficient variability over 50 random splits



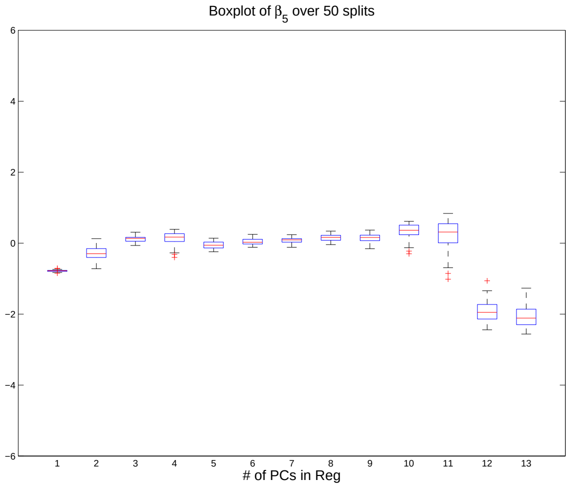
Coefficient variability over 50 random splits



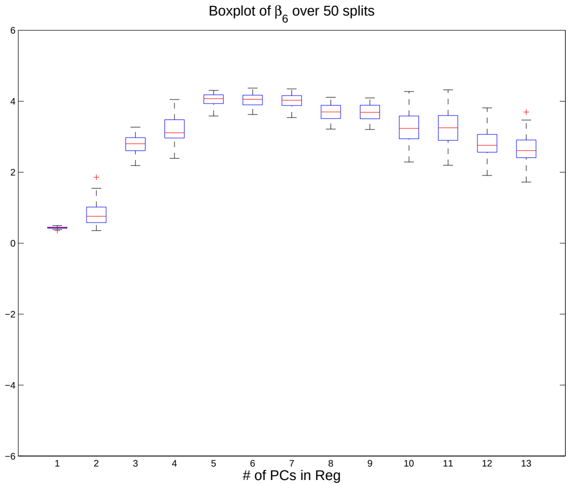
Coefficient variability over 50 random splits



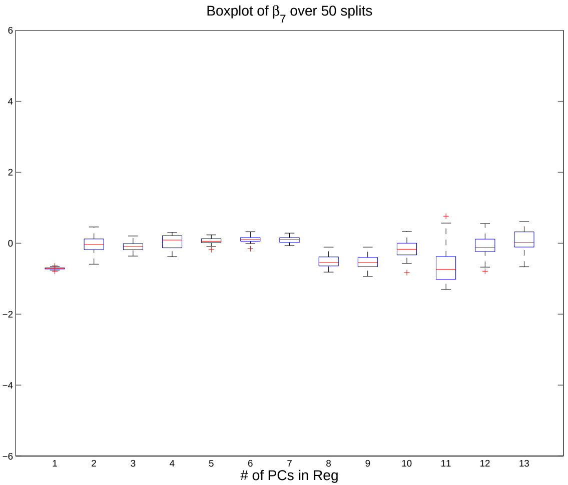
Coefficient variability over 50 random splits



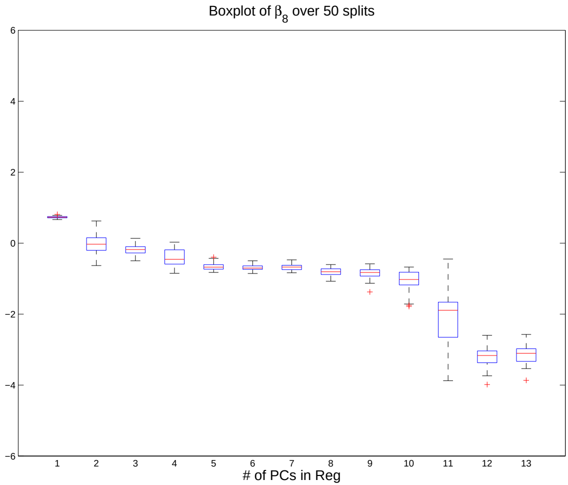
Coefficient variability over 50 random splits



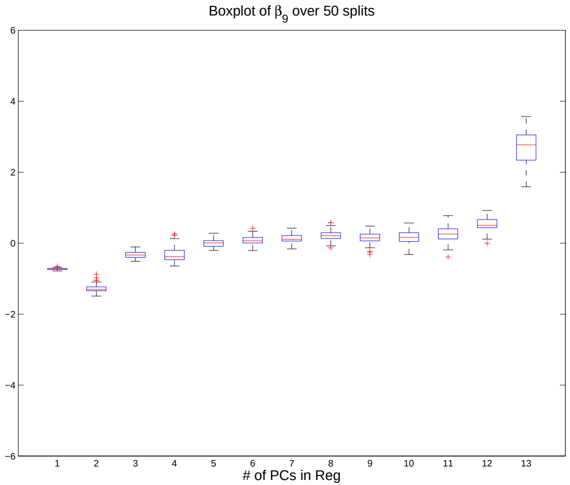
Coefficient variability over 50 random splits



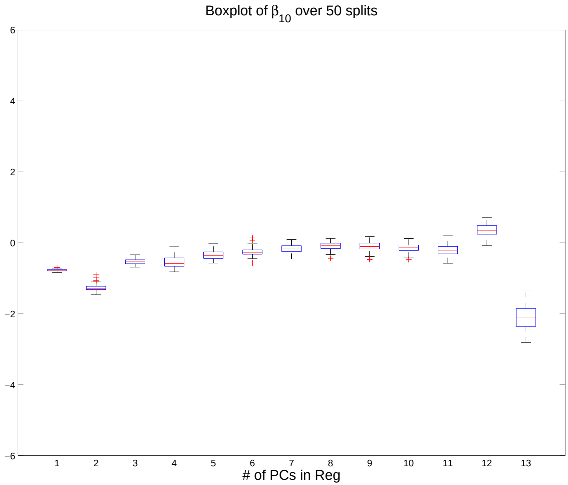
Coefficient variability over 50 random splits



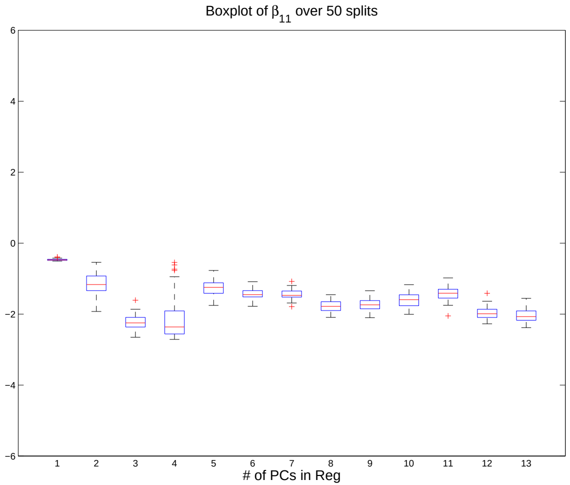
Coefficient variability over 50 random splits



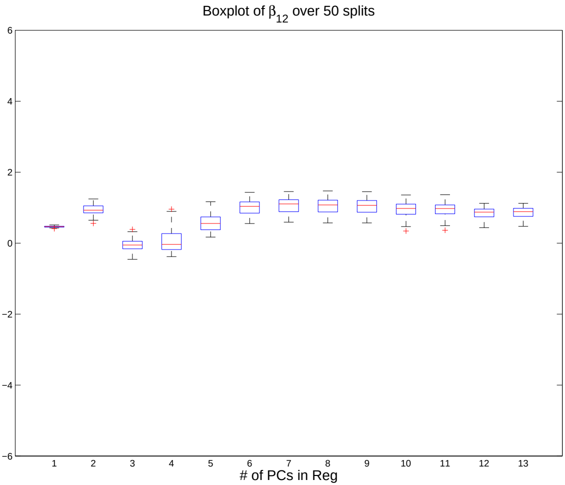
Coefficient variability over 50 random splits



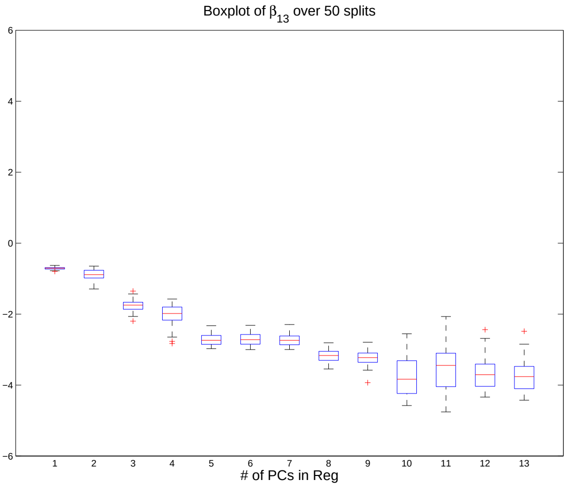
Coefficient variability over 50 random splits



Coefficient variability over 50 random splits



Coefficient variability over 50 random splits



Coefficient variability

- Large variance for X_4, X_{11} and X_{12} with 4 PCs
- Large variance for $X_2, X_3, X_5, X_6, X_7, X_8, X_{13}$ with 11 PCs
- Go back to the whole data (no splitting).
- Check the variance of PCR coefficients:

$$Var(\tilde{\beta}_k) = \sum_{j=1}^{k} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j^T \sigma^2.$$

Coefficient variability

- With $k = 4$ PCs, $\text{diag}(Var(\tilde{\beta}_k)) = \begin{pmatrix} 0.1333 \\ 0.1738 \\ 0.0290 \\ 0.9902 \\ 0.0838 \\ 0.3979 \\ 0.1203 \\ 0.1583 \\ 0.1553 \\ 0.1106 \\ 0.2500 \\ 0.1523 \\ 0.0827 \end{pmatrix}$

PCR summary and related regression method

- Principal Component Regression (PCR) first summarizes multiple explanatory variables into a few principal component directions and then performs regression on those principal component directions.
- These principal component directions are orthogonal to each other, yet contain most of the variations in the explanatory variables.
- Thus, PCR can circumvent the potential numerical difficulty of OLS.
- Partial Least Squares is a related regression technique and it has been widely used in the field of chemometrics.
  - Similar to PCR, PLS also uses a small number of linear transformations of the covariates for regression.
  - PLS makes use of both covariates and the response variable to seek for suitable transformations.